

Towards Grounding Conceptual Spaces in Neural Representations

Lucas Bechberger* and Kai-Uwe Kühnberger

Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany
lucas.bechberger@uni-osnabrueck.de, kai-uwe.kuehnberger@uni-osnabrueck.de

Abstract. The highly influential framework of conceptual spaces provides a geometric way of representing knowledge. It aims at bridging the gap between symbolic and subsymbolic processing. Instances are represented by points in a high-dimensional space and concepts are represented by convex regions in this space. In this paper, we present our approach towards grounding the dimensions of a conceptual space in latent spaces learned by an InfoGAN from unlabeled data.

1 Introduction

The cognitive framework of conceptual spaces [14, 15] attempts to bridge the gap between symbolic and subsymbolic AI by proposing an intermediate conceptual layer based on geometric representations. A conceptual space is a high-dimensional space spanned by a number of quality dimensions representing interpretable features. Convex regions in this space correspond to concepts. Abstract symbols can be grounded by linking them to concepts in a conceptual space whose dimensions are based on subsymbolic representations.

The framework of conceptual spaces has been highly influential in the last 15 years within cognitive science and cognitive linguistics [11, 13, 20]. It has also sparked considerable research in various subfields of artificial intelligence, ranging from robotics and computer vision [5–7] over the semantic web and ontology integration [1, 10] to plausible reasoning [9, 19].

Although this framework provides means for representing concepts, it does not consider the question of how these concepts can be learned from mostly unlabeled data. Moreover, the framework assumes that the dimensions spanning the conceptual space are already given a priori. In practical applications of the framework, they thus often need to be handcrafted by a human expert.

In this paper, we argue that by using neural networks, one can automatically extract the dimensions of a conceptual space from unlabeled data. We propose that latent spaces learned by an InfoGAN [8] (a special class of Generative Adversarial Networks [17]) can serve as domains in the conceptual spaces framework. We further propose to use a clustering algorithm in these latent spaces in order to discover meaningful concepts.

* Corresponding author, ORCID: 0000-0002-1962-1777

The remainder of this paper is structured as follows: Section 2 presents the framework of conceptual spaces and Section 3 introduces the InfoGAN framework. In Section 4, we present our idea of combining these two frameworks. Section 5 gives an illustrative example and Section 6 concludes the paper.

2 Conceptual Spaces

A conceptual space [14] is a high-dimensional space spanned by so-called “quality dimensions”. Each of these dimensions represents an interpretable way in which two stimuli can be judged to be similar or different. Examples for quality dimensions include temperature, weight, time, pitch, and hue. A domain is a set of dimensions that inherently belong together. Different perceptual modalities (like color, shape, or taste) are represented by different domains. The color domain for instance can be represented by the three dimensions hue, saturation, and brightness.¹ Distance within a domain is measured by the Euclidean metric.

The overall conceptual space is defined as the product space of all dimensions. Distance within the overall conceptual space is measured by the Manhattan metric of the intra-domain distances. The similarity of two points in a conceptual space is inversely related to their distance – the closer two instances are in the conceptual space, the more similar they are considered to be.

The framework distinguishes properties like “red”, “round”, and “sweet” from full-fledged concepts like “apple” or “dog”: Properties are represented as regions within individual domains (e.g., color, shape, taste), whereas full-fledged concepts span multiple domains. Reasoning within a conceptual space can be done based on geometric relationships (e.g., betweenness and similarity) and geometric operations (e.g., intersection or projection).

Recently, Balkenius & Gärdenfors [2] have argued that population coding in the human brain can give rise to conceptual spaces. They discuss the connection between neural and conceptual representations from a neuroscience/psychology perspective, whereas we take a machine learning approach in this paper.

3 Representation Learning with InfoGAN

Within the research area of neural networks, there has been some substantial work on learning compressed representations of a given feature space. Bengio et al. [3] provide a thorough overview of different approaches in the representation learning area. They define representation learning as “learning representations of the data that make it easier to extract useful information when building classifiers or other predictors”. We will focus our discussion here on one specific approach that is particularly fitting to our proposal, namely InfoGAN [8]. InfoGAN is an extension of the GAN (Generative Adversarial Networks) framework [17] which has been applied to a variety of problems (e.g., [12, 18, 21–23]). We first describe the original GAN framework before moving on to InfoGAN.

¹ Of course, one can also use other color spaces, e.g., the CIE L*a*b* space.

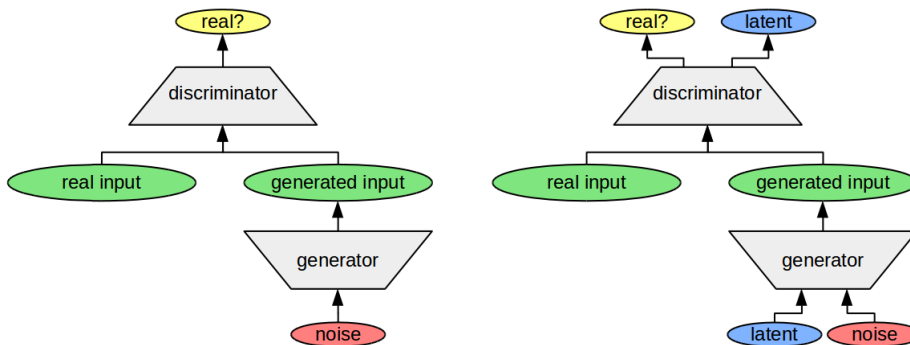


Fig. 1. Left: Illustration of a GAN. Right: Illustration of an InfoGAN.

The GAN framework (depicted in the left part of Figure 1) consists of two networks, the generator and the discriminator. The generator is fed with a low-dimensional vector of noise values. Its task is to create high-dimensional data vectors that have a similar distribution as real data vectors taken from an unlabeled training set. The discriminator receives a data vector that was either created by the generator or taken from the training set. Its task is to distinguish real inputs from generated inputs. Although the discriminator is trained on a classification task, the overall system works in an unsupervised way. The overall architecture can be interpreted as a two-player game: The generator tries to fool the discriminator by creating realistic inputs and the discriminator tries to avoid being fooled by the generator. When the GAN framework converges, the discriminator is expected to make predictions only at chance level and the generator is expected to create realistic data vectors. Although the overall framework works quite well, the dimensions of the input noise vector are usually not interpretable.

Chen et al. [8] have extended the original framework by introducing latent variables: In the InfoGAN framework (shown in the right part of Figure 1), the generator receives an additional input vector. The entries of this vector are values of latent random variables, selected based on some probability distribution that was defined a priori (e.g., uniform or Gaussian). The discriminator has the additional task to reconstruct these latent variables.² Chen et al. argue that this ensures that the mutual information between the latent variable vector and the generated data vector is high. They showed that after training an InfoGAN, the latent variables tend to have an interpretable meaning. For instance, in an experiment on the MNIST data set, the latent variables corresponded to type of digit, digit rotation and stroke thickness. InfoGANs can thus provide a bidirectional mapping between observable data vectors and interpretable latent dimensions: One can both extract interpretable dimensions from a given data vector and create a data vector from an interpretable latent representation.

² This introduces a structure similar to an autoencoder (with the latent variables as input/output and the generated data vector as hidden representation).

4 Using Representation Learning to Ground Domains

For some domains of a conceptual space, a dimensional representation is already available. For instance, the color domain can be represented by the three-dimensional HSB space. For other domains, it is however quite unclear how to represent them based on a handful of dimensions. One prominent example is the shape domain: To the best of our knowledge, there are no widely accepted dimensional models for describing shapes.

We propose to use the InfoGAN framework in order to learn such a dimensional representation based on an unlabeled data set: Each of the latent variables can be interpreted as one dimension of the given domain of interest. For instance, the latent variables learned on a data set of shapes can be interpreted as dimensions of the shape domain. Three important properties of domains in a conceptual space are the following: interpretable dimensions, a distance-based notion of similarity, and a geometric way of describing semantic betweenness. We think that the latent space of an InfoGAN is a good candidate for representing a domain of a conceptual space, because it fulfills all of the above requirements:

As described before, Chen et al. [8] found that the individual latent variables tend to have an interpretable meaning. Although this is only an empirical observation, we expect it generalize to other data sets and thus to other domains.

Moreover, the *smoothness assumption* used in representation learning (cf. [3] and [16, Ch. 15]) states that points with small distance in the input space should also have a small distance in the latent space. This means that a distance-based notion of similarity in the latent space is meaningful.

Finally, Radford et al. [18] found that linear interpolations between points in the latent space of a GAN correspond to a meaningful “morph” between generated images in the input space. This indicates that geometric betweenness in the latent space can represent semantic betweenness.

There are two important hyperparameters to the approach of grounding domains in InfoGANs: The number of latent variables (i.e., the dimensionality of the learned domain) and the type of distribution used for the latent variables (e.g., uniform vs. Gaussian). Note that one would probably aim for the lowest-dimensional representation that still describes the domain sufficiently well.

Finally, we would like to address a critical aspect of this proposal: How can one make sure that the representation learned by the neural network only represents information from the target domain (e.g., shape) and not anything related to other domains (e.g., color)? In our opinion, there are two complementary methods to “steer” the network towards the desired representation:

The first option consists of selecting only such inputs for the training set that do not exhibit major differences with respect to other domains. For instance, a training set for the shape domain should only include images of shapes that have the same color (e.g., black shape on white ground). If there is only very small variance in the data set with respect to other domains (e.g., color), the network is quite unlikely to incorporate this information into its latent representation.

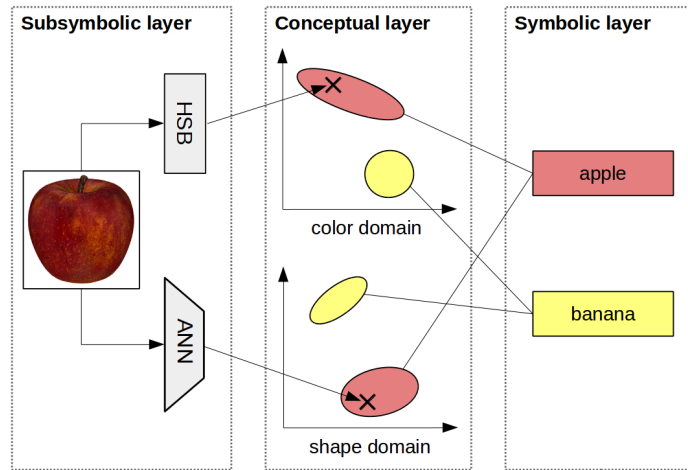


Fig. 2. Illustration of the envisioned overall system. Perceptions from the subsymbolic layer are transformed to points in different domains by feature extractors that are either handcrafted (e.g., HSB for colors) or trained neural networks (e.g., for the shape domain). Each concept is described by one region per domain.

The second option concerns modifications of the network’s loss function: One could for instance introduce an additional term into the loss function which measures the correlation between the learned latent representation and dimensions from other (already defined) domains. This would cause a stronger error signal if the network starts to re-discover already known dimensions from other domains and therefore drive the network away from learning redundant representations.

A simple proof of concept implementation for the shape domain could be based on a data set of simple 2D shapes (circles, triangles, rectangles, etc.) in various orientations and locations. For a more thorough experiment, one could for instance use ShapeNet³ [4], a data base of over 50,000 3D models for more than 50 categories of objects. One could render these 3D models from various perspectives in order to get 2D inputs (for learning to represent 2D shapes) or work on a voxelized 3D input (for learning representations of 3D shapes).

5 An Illustrative Example

Figure 2 illustrates a simplified example of our envisioned overall system. Here, we consider only two domains: color and shape. Color can be represented by the HSB space using the three dimensions hue, saturation and brightness. This is an example for a hard-coded domain. The representation of the shape domain, however, needs to be learned. The artificial neural network depicted in Figure

³ <https://www.shapenet.org/>

2 corresponds to the discriminator of an InfoGAN trained on a data set of shapes.

Let us consider two example concepts: The concept of an apple can be described by the “red” region in the color domain and the “round” region in the shape domain. The concept of a banana can be represented by the “yellow” region in the color domain and the “cylindric” region in the shape domain.

If the system makes a new observation (e.g., an apple as depicted in Figure 2), it will convert this observation into a point in the conceptual space. For the color domain, this is done by a hard-coded conversion to the HSB color space. For the shape domain, the observation is fed into the discriminator and its latent representation is extracted, resulting in the coordinates for the shape domain. Now in order to classify this observation, the system needs to check whether the resulting data point is contained in any of the defined regions. If the data point is an element of the apple region in both domains (which is the case in our example), this observation should be classified as an apple. If the data point is an element of the banana region, the object should be classified as a banana.

Based on a new observation, the existing concepts can also be updated: If the observation was classified as an apple, but it is not close to the center of the apple region in one of the domains, this region might be enlarged or moved a bit, such that the observed instance is better matched by the concept description. If the observation does not match any of the given concepts at all, even a new concept might be created. This means that concepts cannot only be applied for classification, but they can also be learned and updated. Note that this can take place without explicit label information, i.e., in an unsupervised way. Our overall research goal is to develop a clustering algorithm that can take care of incrementally updating the regions in such a conceptual space.

Please note that the updates considered above only concern the connections between the conceptual and the symbolic layer. The connections between the subsymbolic and the conceptual layer remain fixed. The neural network thus only serves as a preprocessing step in our approach: It is trained before the overall system is used and remains unchanged afterwards. Simultaneous updates of both the neural network and the concept description might be desirable, but would probably introduce a great amount of additional complexity.

6 Conclusion and Future Work

In this paper, we outlined how neural representations can be used to ground the domains of a conceptual space in perception. This is especially useful for domains like shape, where handcrafting a dimensional representation is difficult. We argued that the latent representations learned by an InfoGAN have suitable properties for being combined with the conceptual spaces framework. In future work, we will implement the proposed idea by giving a neural grounding to the domain of simple 2D shapes. Furthermore, we will devise a clustering algorithm for discovering and updating conceptual representations in a conceptual space.

References

1. Benjamin Adams and Martin Raubal. Conceptual Space Markup Language (CSML): Towards the Cognitive Semantic Web. *2009 IEEE International Conference on Semantic Computing*, Sep 2009.
2. Christian Balkenius and Peter Gärdenfors. Spaces in the Brain: From Neurons to Meanings. *Frontiers in Psychology*, 7:1820, 2016.
3. Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug 2013.
4. Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. December 2015.
5. Antonio Chella, Haris Dindo, and Ignazio Infantino. Anchoring by Imitation Learning in Conceptual Spaces. *AI*IA 2005: Advances in Artificial Intelligence*, pages 495–506, 2005.
6. Antonio Chella, Marcello Frixione, and Salvatore Gaglio. Conceptual Spaces for Computer Vision Representations. *Artificial Intelligence Review*, 16(2):137–152, 2001.
7. Antonio Chella, Marcello Frixione, and Salvatore Gaglio. Anchoring Symbols to Conceptual Spaces: The Case of Dynamic Scenarios. *Robotics and Autonomous Systems*, 43(2-3):175–188, May 2003.
8. Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2016.
9. Joaquín Derrac and Steven Schockaert. Inducing Semantic Relations from Conceptual Spaces: A Data-Driven Approach to Plausible Reasoning. *Artificial Intelligence*, 228:66–94, Nov 2015.
10. Stefan Dietze and John Domingue. Exploiting Conceptual Spaces for Ontology Integration. In *Data Integration Through Semantic Technology (DIST2008) Workshop at 3rd Asian Semantic Web Conference (ASWC 2008)*, 2008.
11. Igor Douven, Lieven Decock, Richard Dietz, and Paul Égré. Vagueness: A Conceptual Spaces Approach. *Journal of Philosophical Logic*, 42(1):137–160, Nov 2011.
12. Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative Multi-Adversarial Networks. November 2016.
13. Sandro R. Fiorini, Peter Gärdenfors, and Mara Abel. Representing Part-Whole Relations in Conceptual Spaces. *Cognitive Processing*, 15(2):127–142, Oct 2013.
14. Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT press, 2000.
15. Peter Gärdenfors. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press, 2014.
16. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
17. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. June 2014.
18. Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. November 2015.

19. Steven Schockaert and Henri Prade. Interpolation and Extrapolation in Conceptual Spaces: A Case Study in the Music Domain. *Lecture Notes in Computer Science*, pages 217–231, 2011.
20. Massimo Warglien, Peter Gärdenfors, and Matthijs Westera. Event Structure, Conceptual Spaces and the Semantics of Verbs. *Theoretical Linguistics*, 38(3-4), Jan 2012.
21. Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 82–90. Curran Associates, Inc., 2016.
22. Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based Generative Adversarial Network. September 2016.
23. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. March 2017.